RESEARCH PAPER

# Deep Learning with Heterogeneous Graph Embeddings for Mortality Prediction from Electronic Health Records

**Tingyi Wanyan[1,2], Hossein Honarvar[1], Ariful Azad[2], Ying Ding[3,4] & Benjamin S. Glicksberg[1,5†]**

[1]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, New York 10065, USA

[2]School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47405-7000, USA

[3]Dell Medical School, University of Texas at Austin, Austin, Texas 78701-1996, USA

[4]School of Informatics, University of Texas at Austin, Austin, Texas 78712-1139, USA

[5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10065, USA

**Keywords:** Electronic health records (EHRs); Convolutional Neural Networks (CNNs); Heterogeneous Graph Model (HGM); Machine learning; Deep learning

Citation: Wanyan T.Y., et al.: Deep learning with heterogeneous graph embeddings for mortality prediction from electronic health records. Data Intelligence 3(3), 329-339 (2021). doi: 10.1162/dint_a_00097
Received: December 24, 2020; Revised: April 2, 2021; Accepted: April 30, 2021

## ABSTRACT

Computational prediction of in-hospital mortality in the setting of an intensive care unit can help clinical practitioners to guide care and make early decisions for interventions. As clinical data are complex and varied in their structure and components, continued innovation of modelling strategies is required to identify architectures that can best model outcomes. In this work, we trained a Heterogeneous Graph Model (HGM) on electronic health record (EHR) data and used the resulting embedding vector as additional information added to a Convolutional Neural Network (CNN) model for predicting in-hospital mortality. We show that the additional information provided by including time as a vector in the embedding captured the relationships between medical concepts, lab tests, and diagnoses, which enhanced predictive performance. We found that adding HGM to a CNN model increased the mortality prediction accuracy up to 4%. This framework served as a foundation for future experiments involving different EHR data types on important healthcare prediction tasks.

[†] Corresponding author: Benjamin S. Glicksberg (Email: benjamin.glicksberg@mssm.edu; ORCID: 0000-0003-4515-8090).

## 1. INTRODUCTION

Timely prediction of in-hospital mortality within intensive care units (ICU) is beneficial [1, 2] for practitioners to tailor care and allow for earlier interventions to prevent deterioration [3, 4]. Electronic health record (EHR) data consist of information relating to patient encounters with a health system, such as disease diagnoses, vital signs, and medications, among others [5, 6] which are often used for machine learning (ML) predictions for different tasks in the biomedical domain including mortality prediction [7, 8, 9]. The inherent complexity of EHR data often require advanced modeling frameworks to gain robust performance for these tasks. A common modeling approach for EHR research is to use a 2-dimensional convolutional neural networks (CNN) with one dimension as time and the other as clinical features [10, 11, 12]. In healthcare-related CNN models, various medical features are normally concatenated to be directly used as inputs and create embeddings [13, 14, 15]. This form of feature representation can be powerful, but disregards the graphical structure and interconnectivity between medical concepts [16, 17] which can affect the CNN performance especially since EHR data are often sparse due to missingness [10].

In this work, we proposed a Heterogeneous Graph Model (HGM) to create a patient embedding vector, which better accounts for missingness in data for training a CNN model. The HGM model captures the relationships between different medical concept types (e.g., diagnoses and lab tests) due to its graphical structure. This relational representation facilitates capturing more complex patient patterns and encoding similarities.

## 2. METHODOLOGY

### *2.1 Data Set*

We conducted our experiments on de-identified EHR data from MIMIC-III [18]. This data set contains various clinical data relating to patient admission to ICU, such as disease diagnoses in the form of International Classification of Diseases (ICD)-9 codes, and lab test results as detailed in Supplementary Materials. We collected data for 5,956 patients, extracting lab tests every hour from admission. There are a total of 409 unique lab tests and 3,387 unique disease diagnoses observed. The diagnoses were obtained as ICD-9 codes and they were represented using one-hot encoding where one represents patients with disease and zero indicates those without. We binned the lab test events into 6, 12, 24, and 48 hours prior to patient death or discharge from ICU. From these data, we performed mortality predictions that are 10-fold, cross validated.

### *2.2 Convolutional Neural Network Model*

Convolutional neural networks (CNNs) are often used, and perform well, on image processing tasks [19] due to their inherent feature extraction and abstraction ability, which increase the accuracy for classification tasks. There are also studies that have demonstrated encouraging successes in using CNN for EHR analyses. In this work, we used a standard CNN model as the baseline.

Since CNNs typically require two dimensional inputs, we treated time as the horizontal dimension and medical events as the vertical dimension. For the time dimension, we recorded every event with one-hour binned increments with respect to the patient death or discharge time. In this model, the vertical dimension was constructed by concatenating two medical event vectors: lab tests and diagnoses. Every entry of the lab test vector recorded the value of a specific lab test by hour, and we pre-processed the lab test by considering the values between 0.5 and 99.5 percentile to remove any inaccurate measurement. We then normalized the data by calculating the standard score (z-score). We imputed missing lab values with zeros. For the diagnosis vector, the *i*-th entry is 1 if the *i*-th diagnosis is observed; otherwise 0. We treated mortality prediction as a binary classification, for which we used a softmax layer with two dimensions and cross-entropy for loss.

## 2.3 Heterogeneous Graph Model

The features used in baseline CNN model are essentially raw data concatenated together, which do not consider the relationships between medical concepts. We used an HGM to capture these inherent relationships by creating three different types of nodes: patient, lab test, and diagnosis. These different types of nodes are connected by two relation types: tested and diagnosed. These could be represented with two triples:

$$\begin{aligned} &Patient \quad lab : \{patient, tested, lab\} \\ &Patient \quad diagnosis : \{patient, diagnosed, diagnosis\} \end{aligned}$$

The testing relationship shows whether a specific lab test was given to a patient at a specific time, and the diagnosed relationship shows whether a patient was diagnosed with a disease.

To represent the lab test and diagnosis node types, we used multi-hot encoding vector: $X_l \in \{0,1\}^{409}$ and $X_d \in \{0,1\}^{3387}$, and the *i*-th entry with the value of 1 indicating whether a specific lab test was performed or a specific diagnosis was given.

### 2.3.1 Node Embeddings

For capturing the relations between different medical events related to a patient, we first utilized the TransE model to project different types of nodes into the same latent space, and then classified those nodes that were connected as a similar group and the disconnected nodes as a dissimilar group.

The TransE model uses a set of 1) projection matrices and 2) relation vectors. After initialization, projections and translations are optimized end-to-end. Heterogeneous nodes $X_p$, $X_l$, and $X_d$ are projected into a shared latent space with trainable projection matrices $W_p$, $W_i$, and $W_d$ using the nonlinear mappings with Equation (1):

$$\begin{aligned} c_p &= \sigma\left(W_p \cdot X_p\right) \\ c_i^* &= \sigma\left(W_i \cdot X_i\right) \\ c_d^* &= \sigma\left(W_d \cdot X_d\right) \end{aligned} \tag{1}$$

where $\sigma$ is a non-linear activation function and $c_p, c_i^*$, and $c_d^*$ are the latent representations of each type of node. Despite the fact that the EHR uses different dimensions for different data types $X_p$, $X_i$, and $X_d$, all node types are projected into the same latent space. Then we applied translation operations to link these different types of nodes with Equation (2):

$$\begin{aligned} c_i &= c_i^* - r_{ip} \\ c_d &= c_d^* - r_{dp} \end{aligned} \tag{2}$$

where $r_{ip}$ and $r_{dp}$ are the relation vectors connecting patients to lab tests and diagnoses, respectively. $c_i$ and $c_d$ are the semantically translated projection representation into the same latent space of patient embedding $c_p$.

### 2.3.2 Optimization Model

For training the HGM, we applied a skip-gram optimization model, which increases the proximity between embedding points whose corresponding graph nodes are often connected after the projection and translation operations (Equation (3)):

$$\max \sum_{u \in V} \sum_{t \in T_V} log Pr(N_t(u) \mid u) \tag{3}$$

where $N_t(u)$ are the neighborhood vertices of center node $u$, and $t \in T_V$ is the node type. Here, we learned the node embeddings by maximizing the probability of correctly predicting the patient node's associated lab tests and diagnoses. The prediction probability is modeled as a softmax function with Equation (4):

$$Pr(c_t \mid f(u)) = \frac{e^{\vec{c}_t \cdot \vec{u}}}{Z_u} \tag{4}$$

where $\vec{u}$ is the latent representation of patient $u$, $\vec{c}_t$ is the latent representation of lab and diagnosis neighbors of node $u$, and $\vec{c}_t \cdot \vec{u}$ is the inner product of the two embedding vectors representing their similarity. $Z_u$ is the normalization term $Z_u = \sum_{v \in V} e^{\vec{v}_t \cdot \vec{u}}$ that is a sum over all vertices $V$, each of which is represented as $\vec{v}_t$ including all node types. Therefore, Equation (3) is simplified to Equation (5):

$$\mathcal{L}_s = -\sum_{t \in T} \sum_{u \in V} \left[ \sum_{c_t \in N_t(u)} \vec{c_t} \cdot \vec{u} - log Z_u \right] \tag{5}$$

Numerical computation of $Z_u$ is intractable for large-scale graphs. So we adopted a negative sampling strategy to approximate the normalization factor. We eventually used the following optimization function (Equation (6)):

$$\mathcal{L}_s = -\sum_{t \in T} \sum_{u \in V} \left[ \sum_{c_t \in N_t(u)} log \sigma\left(\vec{c_t} \cdot \vec{u}\right) + \sum_{j=1}^{\mathbb{K}} E_{c_j \sim P_v(c_j)} log \sigma\left(-\vec{c_j} \cdot \vec{u}\right) \right] \tag{6}$$

where $\sigma$ $(x)$ is the sigmoid function, which operates on the dot product between $(c_{i^*u})$ and $\sigma(x) = \frac{1}{1+\exp(-x)}$, and $\mathbb{K}$ is the number of negative samples. $P_v(c_j)$ is the negative sampling distribution.

For training HGM, we performed heterogeneous neighborhood sampling by its one-hop connectivity, and picked *Patient* node as the center node, since it has one-hop connections to both *Diagnoses* and

*Lab_test* nodes. Specifically, for one training center *Patient* node, we uniformly sampled 10 *Diagnoses* one-hop direct connected nodes, and 10 *Lab_test* one-hop direct connected nodes. From these sampled 10 *Diagnoses* nodes, we sampled another 10 *Patient* nodes, each having connections with each of the prior 10 *Diagnoses* nodes. In this way, we connected the center patient node with similar other *Patient* nodes by their common diagnoses. We also sampled the patient node which belongs to the next hour corresponding to the center *Patient* node. For negative sampling, we performed uniform sampling through all *Diagnoses* nodes and *Lab_test* nodes that do not have one-hop connections with the center training patient node. We then projected these different nodes into the same latent space through TransE model. After unifying the embeddings for different node types, each concept is represented as a point in a Euclidean space. In this space, we can measure the similarity between any two vectors using dot product.

### 2.3.3 HGM Embeddings with CNN Model

The HGM embedding vector encodes not only a patient's information, but also their relation with diagnoses, lab tests, and subsequent lab test results in time. The patient node is represented as a vector $X_p \in \mathbb{R}^{477}$ containing the numerical values measured from lab tests averaged at that time step. We concatenated the resulting embedding vectors to feed into the baseline CNN vertical feature dimension to form a final feature vector within every hour, and used these new features as the CNN input to predict mortality. In addition, since we encoded time as a relation type, we can infer the embedding vector of time steps with missing data based on information from the previous hour. We visualized this procedure in Figure 1.

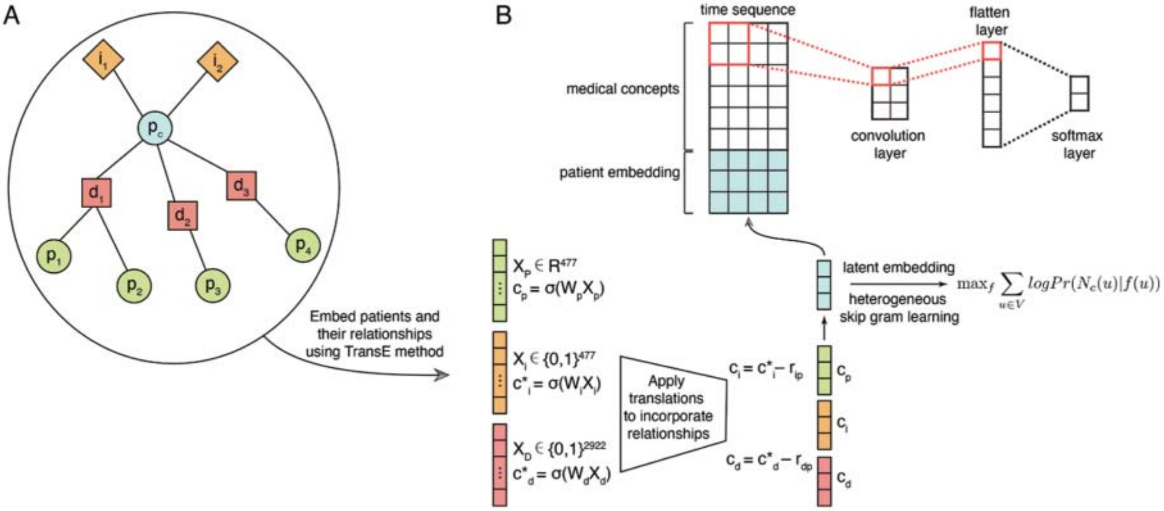


**Figure 1.** (A) A graphical representation of the HGM for p: patient, i: lab test, and d: diagnosis data. (B) All graph nodes in (A) have a corresponding vector like those shown in (B). The vector representations can be projected into a shared space with the TransE method, and this projection is optimized for retaining relations in the original data in the embedding via skip-gram optimization. Finally, these vectors are concatenated into the CNN model for mortality prediction.

## 3. EXPERIMENTS

We aim to predict mortality 6, 12, 24, and 48 hours prior to death and/or discharge. The CNN model was used for prediction as introduced in Section 2.2. The CNN model architecture has two convolutional layers, where the first convolution filter is 5x2, the second layer filter is 3x2, and a maxpooling layer between these two layers. Following the convolution layer is a fully connected layer with latent dimension of 100 neurons. The final layer is a sigmoid layer for predicting the output probability. We compared three different scenarios to test the impact of adding HGM embedding vectors as additional features to the framework:

- HGM: Embed patient labs and diagnosis raw data
- CNN: Use raw lab test feature
- HGM+CNN: Concatenate the HGM patient embedding vector, and the raw lab test feature vector

For baselines, we also compared our developed models with three traditional machine learning models: Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). In this work, we used AUROC and AUPRC scores as the primary performance metric. We tabulated the results in Tables 1 and 2, and we show the evaluation AUROC and AUPRC curves for these tasks in Figure 2.

**Table 1.** Mortality prediction AUROC evaluation.

| Model | Hours prior to death | | | |
|---|---|---|---|---|
| | 6 | 12 | 24 | 48 |
| LG | 0.689±0.01 | 0.691±0.01 | 0.672±0.02 | 0.675±0.02 |
| SVM | 0.654±0.01 | 0.661±0.02 | 0.652±0.01 | 0.653±0.01 |
| RF | 0.667±0.02 | 0.671±0.01 | 0.663±0.02 | 0.654±0.02 |
| HGM | 0.714±0.02 | 0.715±0.03 | 0.653±0.03 | 0.641±0.03 |
| CNN | 0.782±0.01 | 0.771±0.02 | 0.775±0.01 | 0.767±0.01 |
| HGM+CNN | **0.800**±0.01 | **0.791**±0.02 | **0.796**±0.01 | **0.771**±0.01 |

Note: Mean values from 10-fold cross validation with standard deviation for confidence intervals.

**Table 2.** Mortality prediction AUPRC evaluation.

| Model | Hours prior to death | | | |
|---|---|---|---|---|
| | 6 | 12 | 24 | 48 |
| LG | 0.545±0.01 | 0.556±0.02 | 0.542±0.01 | 0.539±0.01 |
| SVM | 0.487±0.02 | 0.501±0.01 | 0.498±0.02 | 0.487±0.02 |
| RF | 0.512±0.02 | 0.523±0.02 | 0.510±0.01 | 0.503±0.01 |
| HGM | 0.557±0.02 | 0.559±0.02 | 0.578±0.02 | 0.567±0.03 |
| CNN | 0.590±0.01 | 0.577±0.02 | 0.589±0.01 | 0.585±0.02 |
| HGM+CNN | **0.601**±0.01 | **0.600**±0.01 | **0.604**±0.01 | **0.617**±0.02 |

Note: Mean values from 10-fold cross validation with standard deviation for confidence intervals.

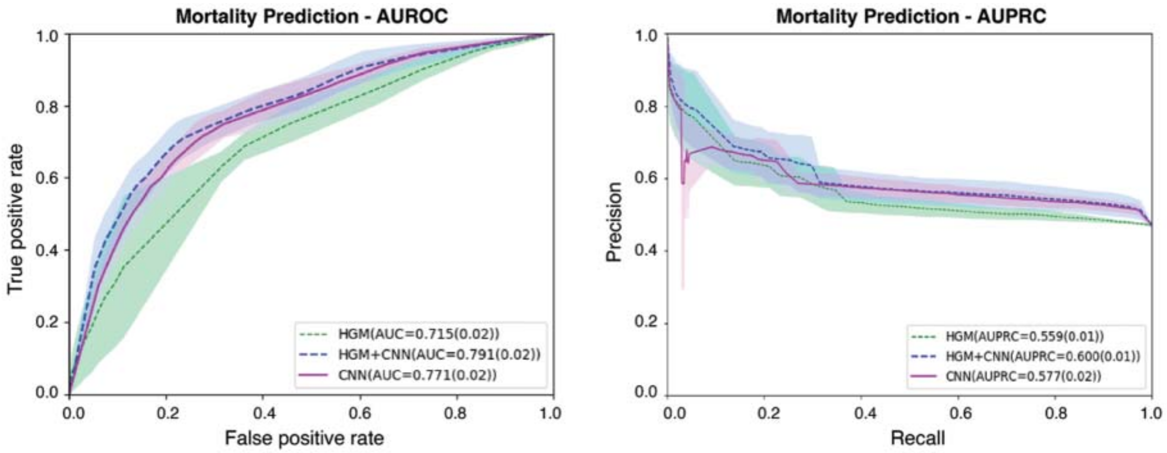


**Figure 2.** Evaluation of AUROC and AUPRC curves for HGM, CNN, and HGM+CNN models.

The testing results show that the HGM+CNN outperforms all baseline models and both the basic HGM and CNN models, indicating the additional information added from the HGM patient embeddings increases the accuracy of predicting in-patient mortality. The prediction accuracy of using different hours prior to death and/or discharge does not vary by much, indicating that different time windows do not have a major impact on the result for this particular task and modeling strategy. The prediction accuracy in the CNN model drops by 1% in the case of six hours prior to death and/or discharge, but not in the other two models, indicating that using the embedding features from HGM model is slightly more robust than the raw data.

## 4. DISCUSSION AND CONCLUSION

In this work, we proposed a method to incorporate patient embedding vector from an HGM model into a CNN model in order to provide more information via interconnectivity between different clinical concepts. We assessed the value of this implementation on a task of predicting mortality in EHR data. The results of our experiment show the superior performance of adding the additional patient embedding vector, which is pretrained from the HGM model, compared to pure raw features as the input to CNN model and traditional ML models, too. In one aspect, this is due to the fact that the HGM embedding vector captures additional relational information between different medical concepts, thus providing additional information to the CNN model.

Furthermore, we observed that concatenating the HGM embedding vector with diagnosis feature vectors did not increase the accuracy *versus* using the concatenation between raw lab test and diagnosis feature vectors. This finding indicates that the raw lab test feature vector can provide unique information for CNN to utilize. At the same time, this finding indicates that the embedded patient vector from HGM model could lose some information from the raw lab test feature along the process of projecting these data into a low dimensional latent space. By concatenating all feature vectors, we aim to preserve the information from different data points, which helps to achieve higher mortality prediction accuracy. There are a few limitations to this study. First, these findings need to be replicated in another data set. Also, exploring more baselines other than the ones shown in this work is beneficial for evaluating the improvements overall. We hope the findings from this work can be expanded in future directions that may add more EHR node types and time components on a variety of other important health-related predictive tasks.

## AUTHOR CONTRIBUTIONS

T.Y. Wanyang (tingyi.wanyan@mssm.edu), A. Azad (azad@iu.edu), Y. Ding (ying.ding@austin.utexas.edu), and B.S. Glicksberg (benjamin.glicksberg@mssm.edu) conceived of the project. T.Y. Wanyang and B.S. Glicksberg collected the data. TW and H. Honarvar (hoseinhonarvar@gmail.com) performed the analyses and made the figures. T.Y. Wanyang, HH, and B.S. Glicksberg wrote the manuscript. TW, HH, A. Azad, Y. Ding, and B.S. Glicksberg edited the manuscript and provided revisions. A. Azad, Y. Ding, and B.S. Glicksberg jointly supervised the work.

## DATA AVAILABILITY STATEMENT

All the data are available in the Science Data Bank repository, https://doi.org/10.11922/sciencedb.j00104.00094, under an Attribution 4.0 International (CC BY 4.0).

## REFERENCES

[1] Johnson, A.E.W., Mark, R.G.: Real-time mortality prediction in the Intensive Care Unit. In: AMIA Annual Symposium Proceedings, pp. 1–10 (2017)

[2] Sharma, A., et al.: Mortality prediction of ICU patients using Machine Leaning: A survey. In: Proceedings of the International Conference on Compute and Data Analysis, pp. 49–53 (2017)

[3] Delahanty, R.J., et al.: Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. Annals of Emergency Medicine 73 (4), 334–344 (2019)

[4] Meyer, A., et al.: Machine learning for real-time prediction of complications in critical care: A retrospective study. The Lancet Respiratory Medicine 6(12), 905–914 (2018)

[5] Glicksberg, B.S., Johnson, K.W., Dudley, J.D.: The next generation of precision medicine: Observational studies, electronic health records, biobanks and continuous monitoring. Human Molecular Genetics 27, R1, R56–R62 (2018)

[6] Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: Towards better research applications and clinical care. Nature Reviews Genetics 13(6), 395–405 (2012)

[7] Glicksberg, B.S., et al.: Automated disease cohort selection using word embeddings from Electronic Health Records. In: Pacific Symposium on Biocomputing 2018, pp. 145-156 (2018)

[8] Rajkomar, A., et al.: Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine 1(1), 18 (2018)

[9] Shickel, B., et al.: Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics 22(5), 1589–1604 (2017)

[10] Cheng, Y., et al.: Risk prediction with electronic health records: A deep learning approach. In: Proceedings of the 2016 SIAM International Conference on Data Mining, pp. 432–440 (2016)

[11] Kim, S.Y., et al.: A deep learning model for real-time mortality prediction in critically ill children. Critical Care 23(1), 279 (2019)

[12] Zhang, J., Gong, J., Barnes, L.: HCNN: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records. In: 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 214–221 (2017)

[13] De Freitas, J.K., et al.: Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. medRxiv preprint medRxiv: 10.1101/2020.11.14.20231894 (2020)

[14] Landi, I., et al.: Deep representation learning of electronic health records to unlock patient stratification at scale. Digital Medicine 3(1), 96 (2020)

[15] Miotto, R., et al.: Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports 6(1) 1–10 (2016)

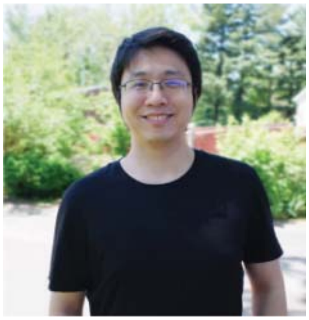
[16] Choi, E., et al.: Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), pp. 4552–4562 (2018)

[17] Choi, E., et al.: Graph convolutional transformer: Learning the graphical structure of electronic health records. arXiv preprint arXiv:1906.04716 (2019)

[18] Johnson, A.E.W., et al.: MIMIC-III, a freely accessible critical care database. Scientific Data 3(1), 1–9 (2016)

[19] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM 60(6), 84–90 (2017)

## AUTHOR BIOGRAPHY



**Tingyi Wanyan** is a PhD student in Artificial Intelligence from Indiana University co-mentored by Dr. Ariful Azad from Indiana University, and Dr. Ying Ding from University of Texas at Austin. He is a visiting PhD student in the Glicksberg Lab at the Icahn School of Medicine at Mount Sinai. He is interested in research involving AI implemented in clinical care, especially regarding integrating various modalities of clinical data via representation learning through Heterogeneous Knowledge Graph models. He specializes in integrating various data types such as electronic health records, medical images, and clinical note.
ORCID: 0000-0002-5011-3973

**Hossein Honarvar** is a postdoctoral fellow at the Hasso Plattner Institute for Digital Health at Mount Sinai in New York City. Previously, he was a postdoctoral researcher in Computational Physics at JILA Research Institute in Boulder and he received his PhD from the University of Colorado Boulder in 2018. His current research focuses on developing interpretable, fair, and multimodal deep learning models for electrocardiogram data and creating novel clinical applications.
ORCID: 0000-0002-1592-2759

**Ariful Azad** is an Assistant Professor of Intelligent Systems Engineering at Luddy School of Informatics, Computing, and Engineering in Indiana University. Dr. Azad obtained his PhD from Purdue University and B.S. from Bangladesh University of Engineering and Technology, Bangladesh. His research interests are in graph machine learning, sparse matrix algorithms, high-performance computing, and bioinformatics.
ORCID: 0000-0003-1332-8630

**Ying Ding** is Bill & Lewis Suit Professor at School of Information, University of Texas at Austin. Before that, she was a professor and director of graduate studies for data science program at School of Informatics, Computing, and Engineering at Indiana University. She has led the effort to develop the online data science graduate program for Indiana University. She also worked as a senior researcher at Department of Computer Science, University of Innsburck (Austria) and Free University of Amsterdam (The Netherlands). She has been involved in various NIH, NSF and European-Union funded projects. She has published more than 240 papers in journals, conferences, and workshops, and served as the program committee member for over 200 international conferences. She is the co-editor of book series called *Semantic Web Synthesis* by Morgan & Claypool publisher, the co-editor-in-chief for *Data Intelligence* published by MIT Press and Chinese Academy of Sciences, and serves as the editorial board member for several top journals in Information Science and Semantic Web. She is the co-founder of Data2Discovery company advancing cutting edge AI technologies in drug discovery and healthcare. Her current research interests include data-driven science of science, AI in healthcare, Semantic Web, knowledge graph, data science, scholarly communication, and the application of Web technologies.
ORCID: 0000-0003-2567-2009

**Benjamin Glicksberg** is an Assistant Professor of Genetics and Genomic Sciences and a member of the Hasso Plattner Institute for Digital Health at the Icahn School of Medicine at Mount Sinai, New York. Dr. Glicksberg has extensive experience in clinical informatics and work involving electronic health record data. He uses machine learning to couple multi-omic patient health data to forward personalized medicine. He completed his PhD in Neuroscience at the Icahn School of Medicine at Mount Sinai in 2017 and post-doctoral work at the University of California, San Francisco in 2019.
ORCID: 0000-0003-4515-8090